Exploration and Modeling of Attrition Data

Daniel Kadyrov

May 8, 2020

1 Introduction

An attrition dataset was used to predict if an employee is active or terminated based on features like age, sex, and education levels. Modeling was performed using a Random Forest classifier and a Support Vector Machine.

2 Data Preprocessing

Initial data features columns including annual and hourly rates, ethnicity, age, sex, job group, first job, and education level. Columns like employee id, termination year, job code, and referral source were removed because they had missing data or unnecessary data for the classification. Status, whether the individual is active or terminated, was selected as the target column and factorized.

Table 1: Data before Processing

Feature	Value

2.1 Factorization

Features with a wide range or categorical data needed to be factorized. Annual rate was split based on \$20,000, \$50,000, \$75,000, \$100,000, and \$2,000,000. Hourly rate was split based on \$25, \$50, \$75, \$100, and \$1000. Age was split based on 20, 30, 40, 50, 60, 100. Hire month was split into quarters of a year, Q1, Q2, Q3, and Q4. Ethnicity, sex, marital status, number of teams, first job, travel requirements, disabled, veteran, job group, and education were factorized.

Table 2: Data after Processing

Feature	Value

2.2 Exploration of Data

There were 21 total number of features with 9612 rows of data. Examining the correlations between the different features with their affect on the status of the employee.

Table 3: Feature Correlation

Feature 1	Feature 2

3 Modeling

3.1 Feature Selection

Feature selection was performed using Recursive Feature Elimination with a logistic regression model. The following 10 features were selected by the algorithm.

Table 4: Feature Selection

Selected Features

3.2 Training Test Split

The data was split into 70% training and 30% test subsets.

3.3 Random Forest

Random Forest classification was performed on the training data and the model was used to predict the test data for accuracy comparison. The classification report for the Random Forest model.

Table 5: RF Classification Report

Metric	Value

3.4 Support Vector Machine

Support Vector Machine classification was performed on the training data and the model was used to predict the test data for accuracy comparison. The classification report for the SVM model.

Table 6: SVM Classification Report

Metric	Value

3.5 Model Accuracy

The accuracy score of the models was performed based on their predictions of the test data.

Table 7: Model Accuracy

Model	Accuracy

4 Conclusion

Random Forest classification and Support Vector Machine were used to predict the attrition of employees based on features in a dataset. The SVM performed better than the RF classifier.